VideoCapsuleNet: A Simplified Network for Action Detection

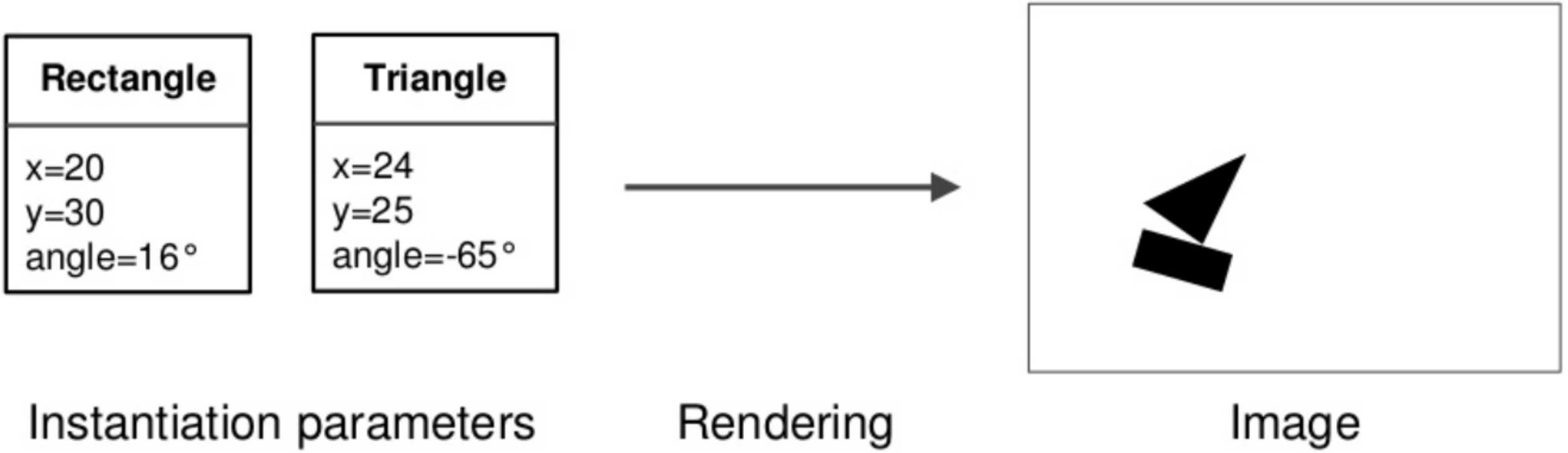
Kevin Duarte, Yogesh S Rawat, and Mubarak Shah

Overview of Capsule Networks

Motivation:

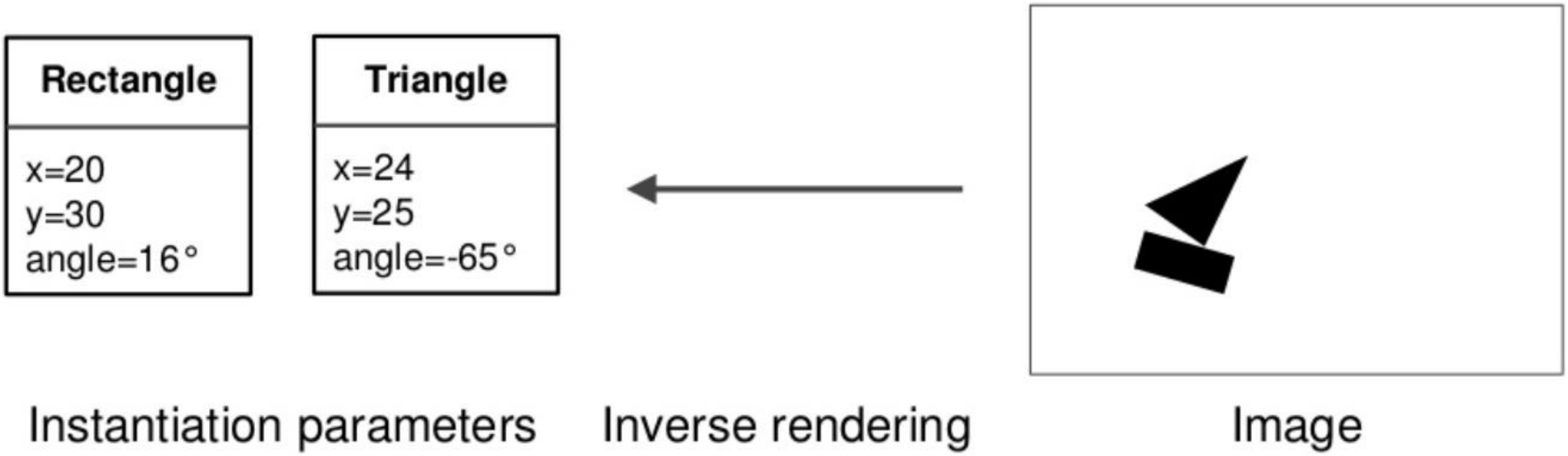
- CNNs do not explicitly model entities
- Add extra structure to CNNs to model entities
  - Entities modeled using a group of neurons
  - Routing-by-agreement to model part-to-whole relationships
- Capsules take inspiration from Inverse Graphics

Computer Graphics



Aurélien Géron (2017). Introduction to Capsule Networks (CapsNets).
https://www.slideshare.net/aureliengeron/introduction-to-capsule-networks-capsnets

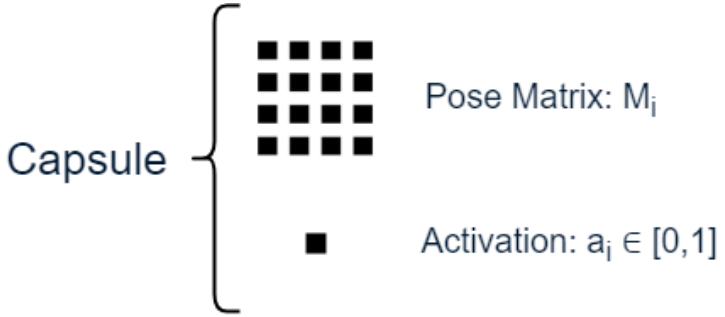
Inverse Graphics



Aurélien Géron (2017). Introduction to Capsule Networks (CapsNets).
https://www.slideshare.net/aureliengeron/introduction-to-capsule-networks-capsnets

Capsules

- 4x4 pose matrix and an activation neuron



- Capsules in lower levels “vote” on capsules in higher level
- Vote V_{ij} is given by:
  - $V_{ij} = M_i W_{ij}$
- Vote is parts’ prediction of the whole
- Votes are grouped together using “Routing by Agreement”

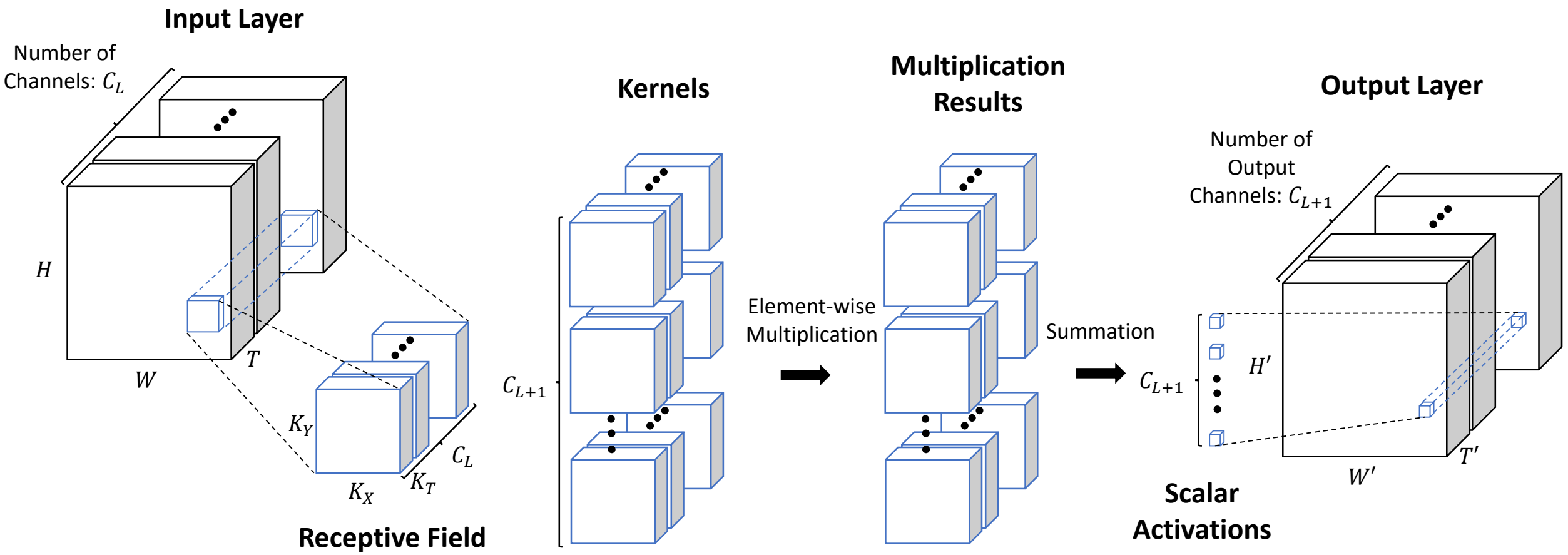
Capsule Networks

- Achieves good results classifying small images (MNIST and smallNorb)
- Has not been successfully applied on high dimensional data
  - Large images or videos
- Issues:
  - Computationally costly
  - Deeper networks cannot fit into memory

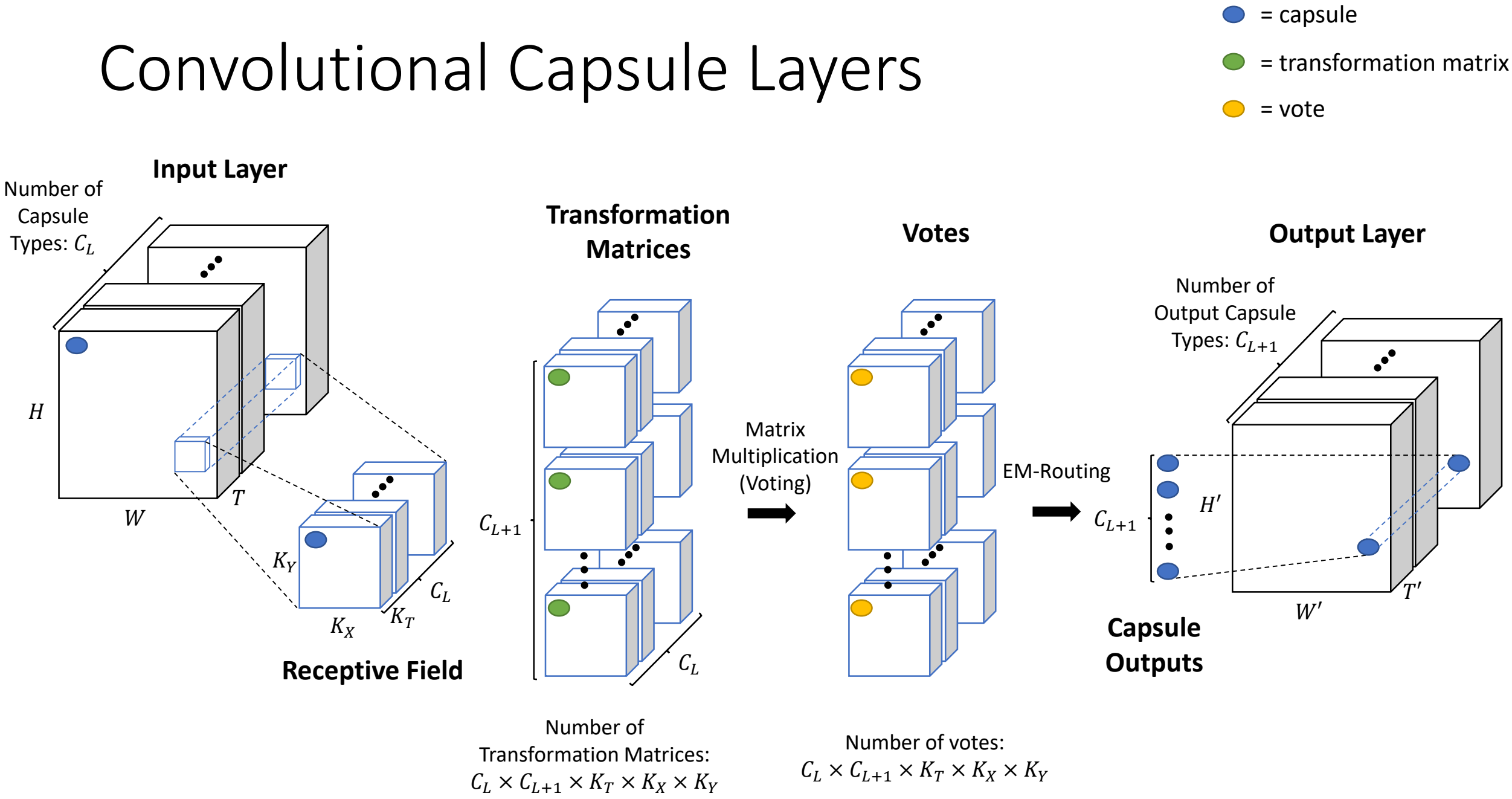
Computational Cost of Capsule Voting

- Convolutional Capsule Routing
  - Routing requires $C_L \times C_{L+1} \times K_T \times K_X \times K_Y$ votes
  - Computed at each pixel
- Number of votes becomes too large if
  - the size capsule layer dimensions is large
  - the size of the receptive field is large

Conventional Convolutional Layers



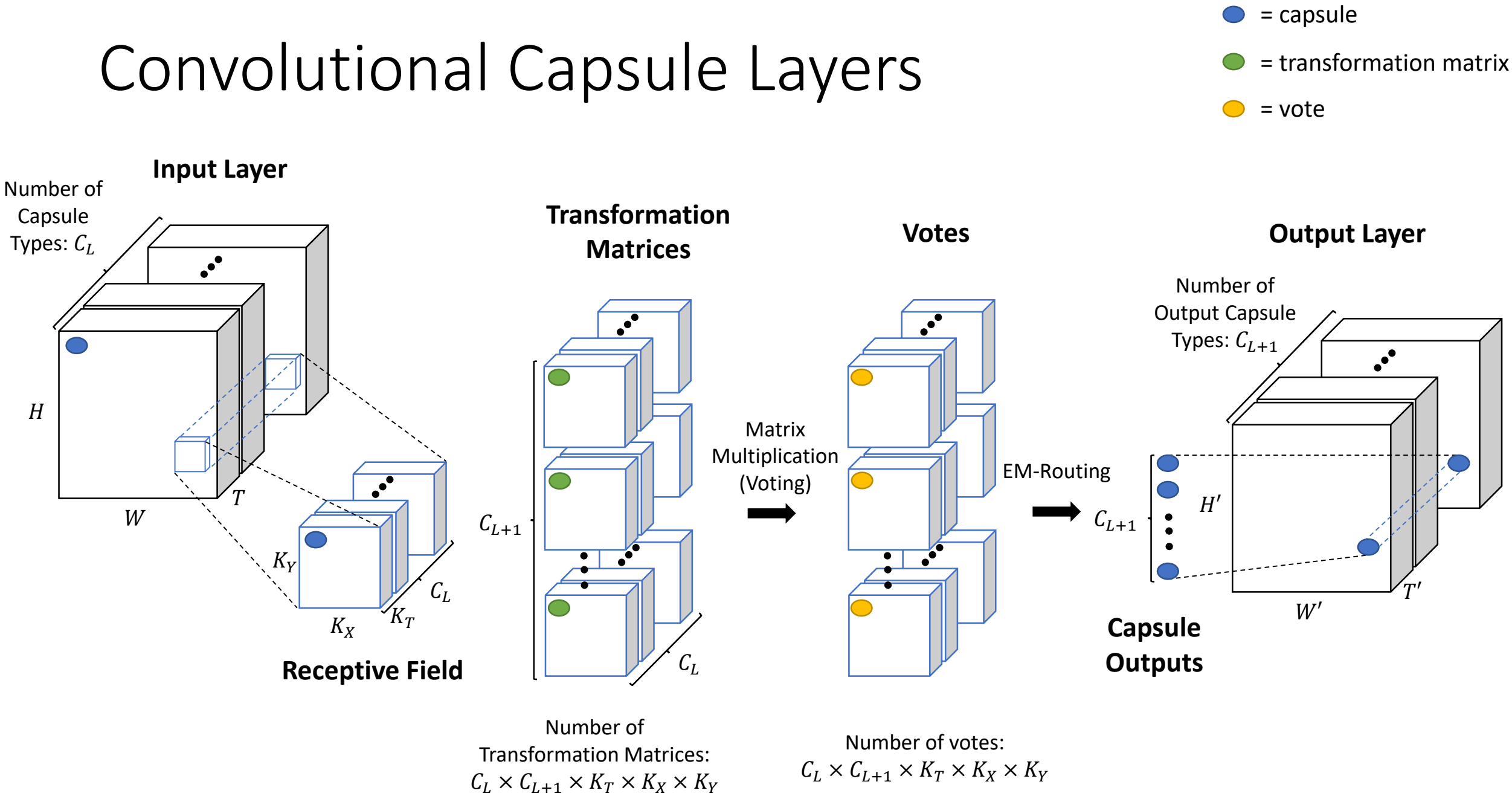
Convolutional Capsule Layers



Two Simplification

- Same transformation matrix for capsules of the same type
- Capsule Pooling

Convolutional Capsule Layers

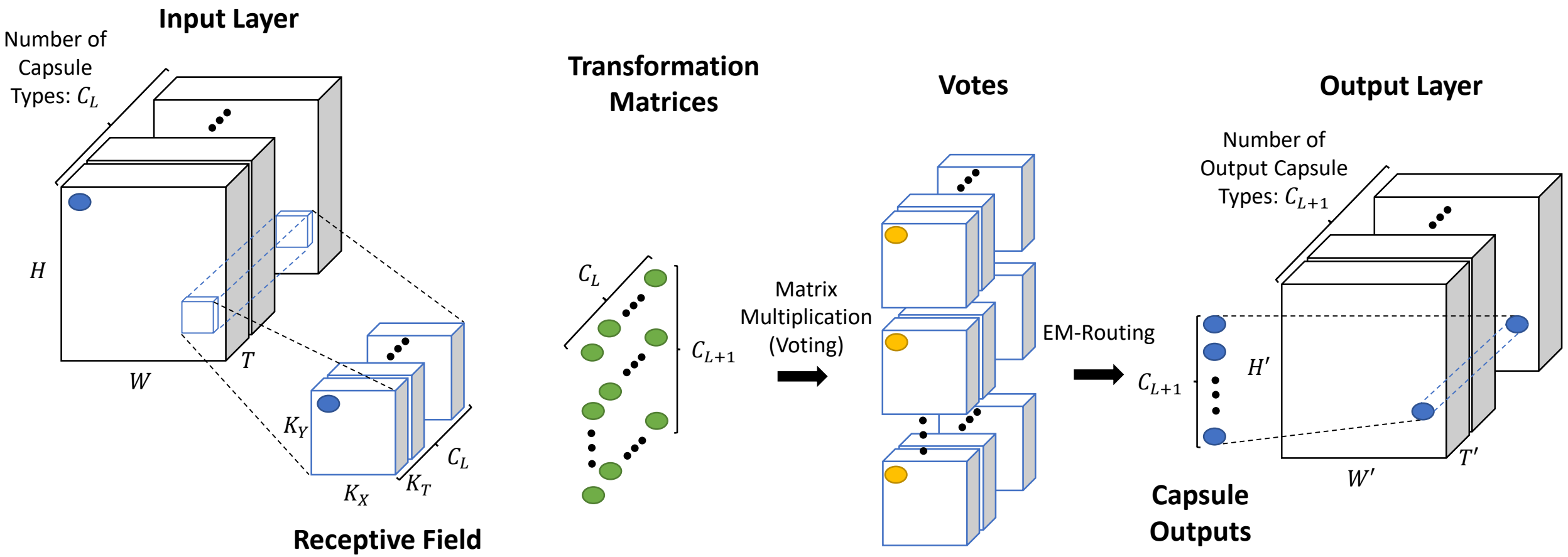


Convolutional Capsule Layers

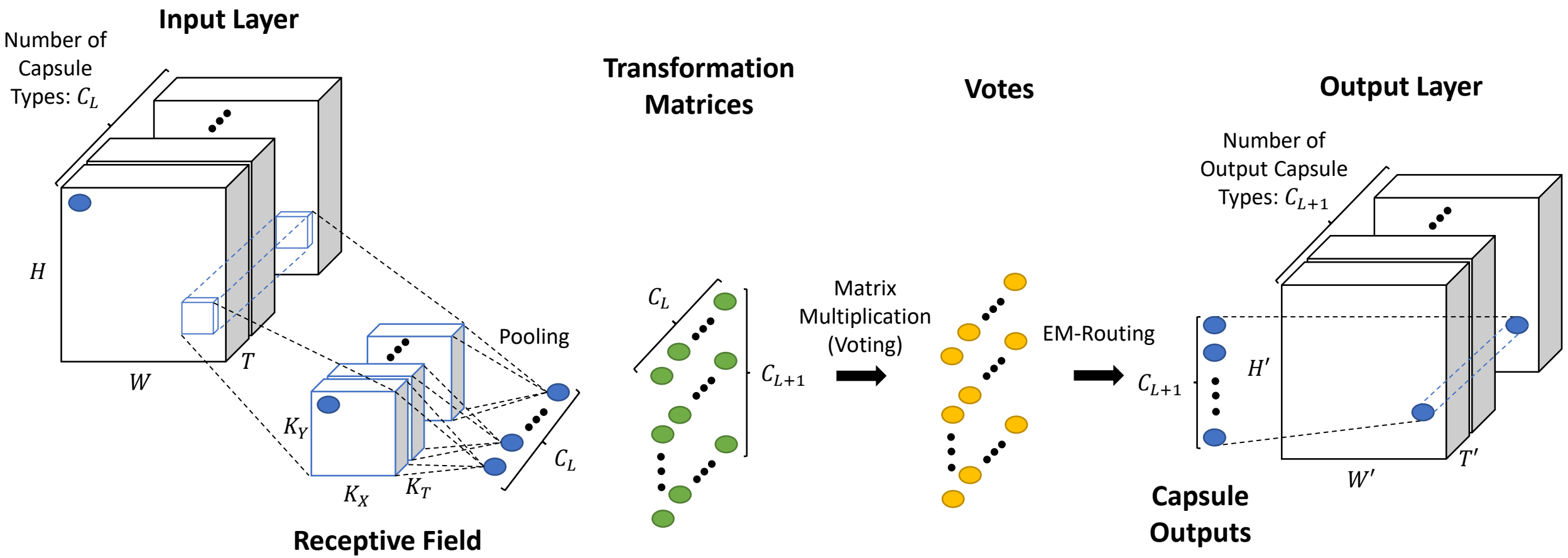
= capsule

= transformation matrix

= vote



Capsule Pooling



Current Video Action Detection Networks

- Require complex multi-stage pipeline:
  - Use Region Proposal network
  - Then classify these regions and perform bounding-box regressions
- Often require optical flow
- Networks are rarely trained end-to-end

VideoCapsuleNet Architecture

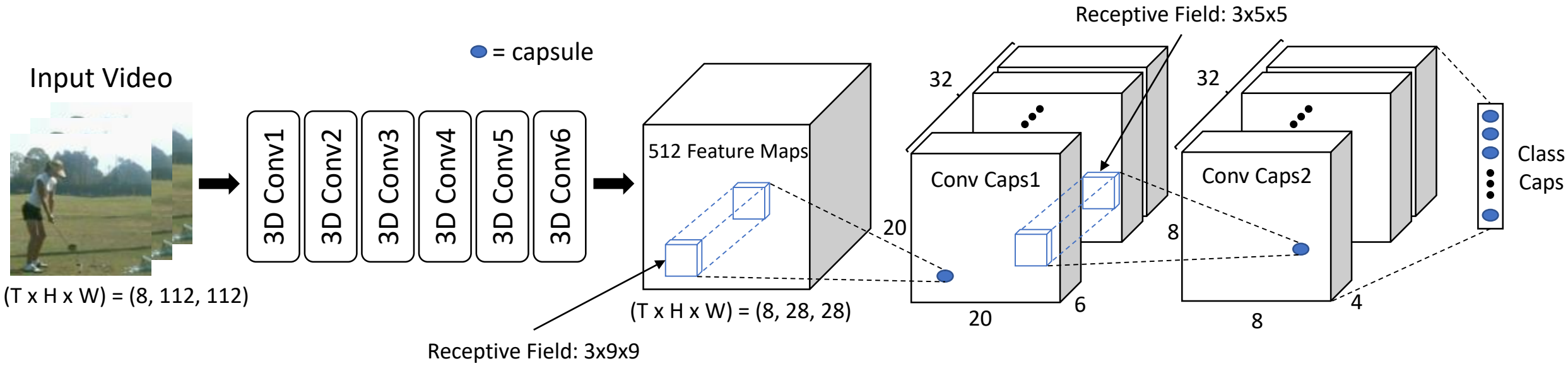
Input Video

(T x H x W) = (8, 112, 112)

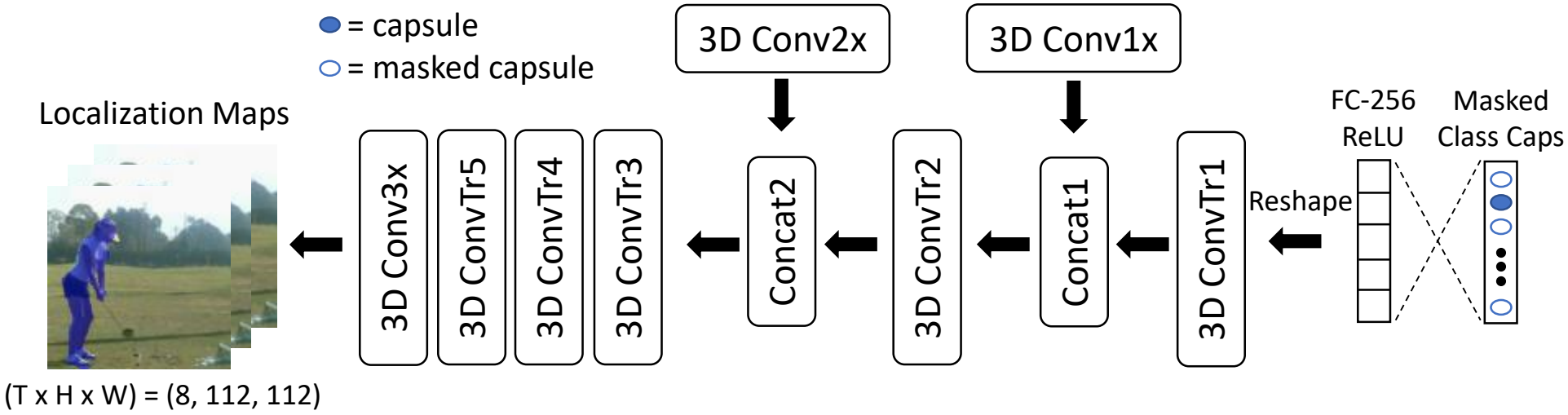
Localization Maps

(T x H x W) = (8, 112, 112)

Encoder


Input Video
(T x H x W) = (8, 112, 112)
3D Conv1
3D Conv2
3D Conv3
3D Conv4
3D Conv5
3D Conv6
= capsule
512 Feature Maps
(T x H x W) = (8, 28, 28)
Receptive Field: 3x9x9
Receptive Field: 3x5x5
32
Conv Caps1
20
20
6
32
Conv Caps2
8
8
4
Class Caps

Decoder


= capsule
= masked capsule
Localization Maps
(T x H x W) = (8, 112, 112)
3D Conv3x
3D ConvTr5
3D ConvTr4
3D ConvTr3
Concat2
3D Conv2x
3D ConvTr2
Concat1
3D Conv1x
3D ConvTr1
Reshape
FC-256
ReLU
Masked
Class Caps

Coordinate Addition

- Add the position (T, H, W) of the capsule to the vote matrix

Capsule at position
(T, H, W) = (1, 4, 2)

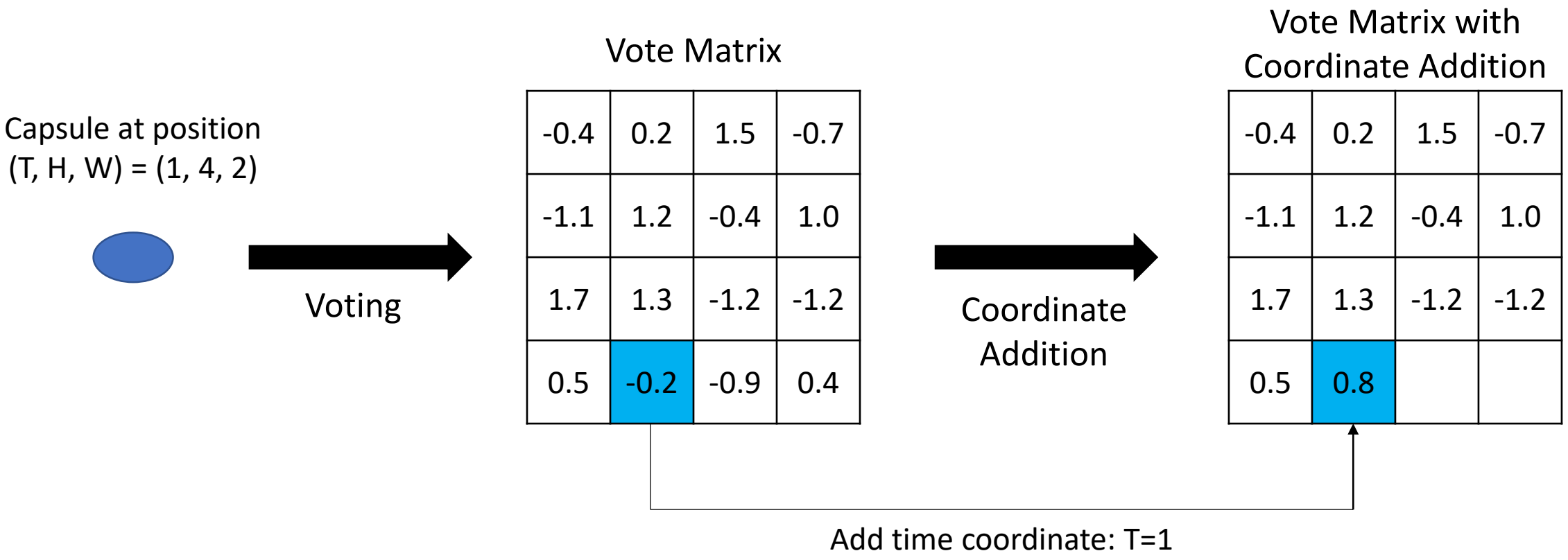
Voting

Vote Matrix

-0.4	0.2	1.5	-0.7
-1.1	1.2	-0.4	1.0
1.7	1.3	-1.2	-1.2
0.5	-0.2	-0.9	0.4

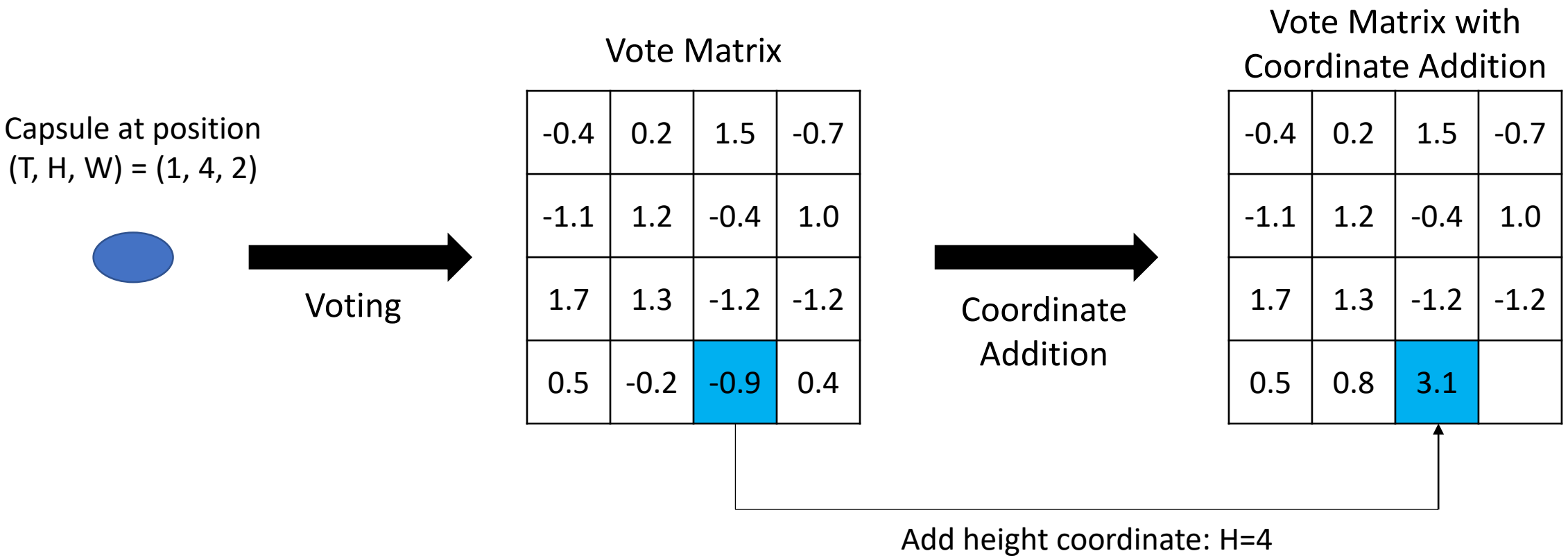
Coordinate Addition

- Add the position (T, H, W) of the capsule to the vote matrix



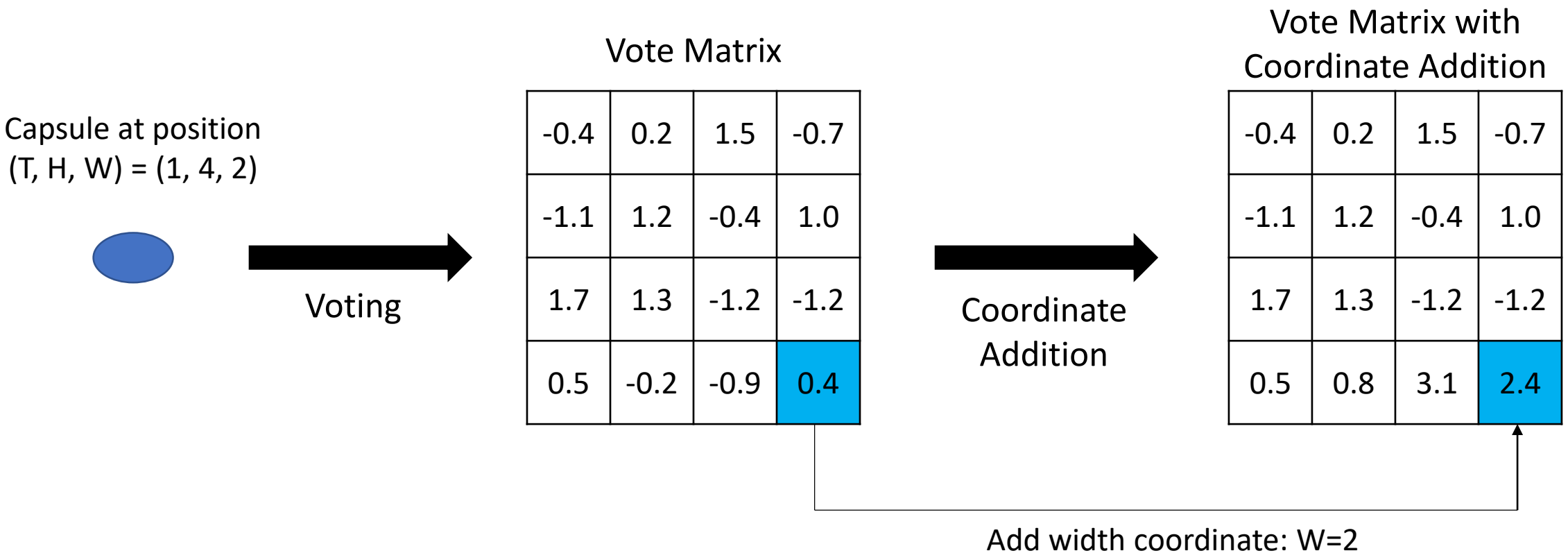
Coordinate Addition

- Add the position (T, H, W) of the capsule to the vote matrix



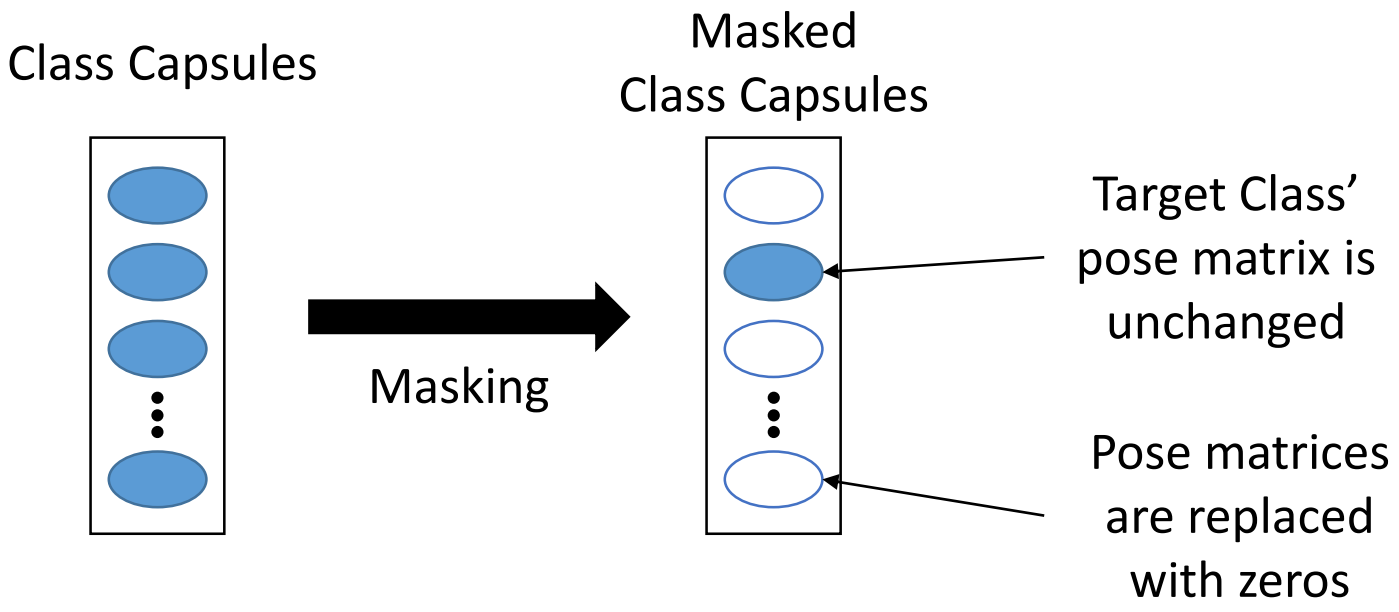
Coordinate Addition

- Add the position (T, H, W) of the capsule to the vote matrix



Capsule Masking

- Training: zero out all class capsules except for ground truth class
- Testing: zero out all class capsules except predicted class



VideoCapsuleNet Training

- Trained end-to-end using the sum of two losses
- Classification Loss:
  - $L_C = \sum_{i \neq t} \max\left(0, m - (a_t - a_i)\right)^2$
- Segmentation Loss:
  - $L_S = -\frac{1}{TXY} \sum_{k=1}^{T} \sum_{i=1}^{X} \sum_{j=1}^{Y} \left[\hat{p}_{kij} \log p_{kij} + \left(1 - \hat{p}_{kij}\right) \log\left(1 - p_{kij}\right)\right]$
- Total Loss:
  - $L = L_C + \lambda L_S$

Action Localization Accuracy

Method
Saha et al. [18]
Peng et al. [8]
Singh et al. [19]
Kalogeiton et al. [4]
Hou et al. [2]
Gu et al. [3]
He et al. [20]
VideoCapsuleNet

Action Localization Accuracy

Method	UCF-Sports	
	f-mAP	v-mAP
	0.5	0.2
Saha et al. [18]	-	-
Peng et al. [8]	84.5	94.8
Singh et al. [19]	-	-
Kalogeiton et al. [4]	87.7	92.7
Hou et al. [2]	86.7	95.2
Gu et al. [3]	-	-
He et al. [20]	-	96.0
VideoCapsuleNet	83.9	**97.1**

Action Localization Accuracy

Method	UCF-Sports		J-HMDB	
	f-mAP	v-mAP	f-mAP	v-mAP
	0.5	0.2	0.5	0.2
Saha et al. [18]	-	-	-	72.6
Peng et al. [8]	84.5	94.8	58.5	74.3
Singh et al. [19]	-	-	-	73.8
Kalogeiton et al. [4]	87.7	92.7	65.7	74.2
Hou et al. [2]	86.7	95.2	61.3	78.4
Gu et al. [3]	-	-	73.3	-
He et al. [20]	-	96.0	-	79.7
VideoCapsuleNet	83.9	**97.1**	64.6	**95.1**

Action Localization Accuracy

Method	UCF-Sports		J-HMDB		UCF-101				
	f-mAP	v-mAP	f-mAP	v-mAP	f-mAP	v-mAP			
	0.5	0.2	0.5	0.2	0.5	0.1	0.2	0.3	0.5
Saha et al. [18]	-	-	-	72.6	-	76.6	66.8	55.5	35.9
Peng et al. [8]	84.5	94.8	58.5	74.3	65.7	77.3	72.9	65.7	35.9
Singh et al. [19]	-	-	-	73.8	-	-	73.5	-	46.3
Kalogeiton et al. [4]	87.7	92.7	65.7	74.2	69.5	-	77.2	-	51.4
Hou et al. [2]	86.7	95.2	61.3	78.4	67.3	77.9	73.1	69.4	-
Gu et al. [3]	-	-	73.3	-	76.3	-	-	-	59.9
He et al. [20]	-	96.0	-	79.7	-	-	71.7	-	-
VideoCapsuleNet	83.9	**97.1**	64.6	**95.1**	**78.6**	**98.6**	**97.1**	**93.7**	**80.3**

Qualitative Results – Entire Videos

Cricket Bowling

Biking

Diving

Floor Gymnastics

Effects of Capsule Masking

- Allows for class specific localizations
  - Giving the ground truth target class results in better localization accuracies

Target Class	UCF-Sports		J-HMDB		UCF-101	
	f-mAP	v-mAP	f-mAP	v-mAP	f-mAP	v-mAP
Predicted Class	83.9	97.1	64.6	95.1	78.6	80.3

Effects of Capsule Masking

- Allows for class specific localizations
  - Giving the ground truth target class results in better localization accuracies

Target Class	UCF-Sports		J-HMDB		UCF-101	
	f-mAP	v-mAP	f-mAP	v-mAP	f-mAP	v-mAP
Predicted Class	**83.9**	97.1	64.6	95.1	78.6	80.3
Ground Truth Class	82.8	97.1	**66.8**	**95.4**	**80.1**	**82.0**

Ablations: Coordinate Addition

- Improves VideoCapsuleNet's classification and localization accuracy

	Without Coordinate Addition	With Coordinate Addition
Accuracy	71.7	**79.0**
f-mAP	72.9	**78.6**
v-mAP	74.9	**80.3**

Ablations: Extra Skip Connections

- We add skip connections at convolutional layers
  - Increases number of network parameters
  - Network speed is greatly decreased

	With Extra Skip Connections	Without Extra Skip Connections
Accuracy	78.7	**79.0**
f-mAP	77.4	**78.6**
v-mAP	**80.7**	80.3

Ablations: # of Convolutional Layers

- We vary the number of 3D-convolutional layers prior to the capsules

	4 Conv Layers	6 Conv Layers	8 Conv Layers
Accuracy	74.6	**79.0**	71.4
f-mAP	72.1	**78.6**	70.4
v-mAP	73.5	**80.3**	71.3

Ablations: Losses and Reconstruction

	L_C	L_S	$L_C + L_s$
Accuracy	62.0	-	**79.0**
f-mAP	-	51.1	**78.6**
v-mAP	-	48.1	**80.3**

Ablations: Losses and Reconstruction

- We include the reconstruction loss L_r in our end-to-end training

	L_C	L_S	$L_C + L_r$	$L_C + L_s$
Accuracy	62.0	-	72.2	**79.0**
f-mAP	-	51.1	-	**78.6**
v-mAP	-	48.1	-	**80.3**

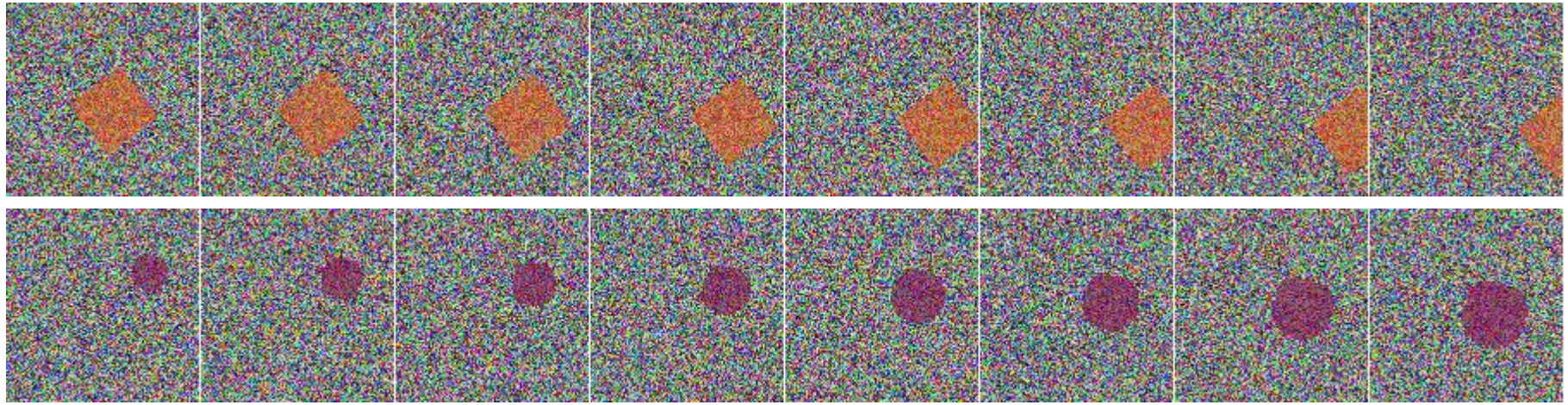
Ablations: Losses and Reconstruction

- We include the reconstruction loss L_r in our end-to-end training

	L_C	L_S	$L_C + L_r$	$L_C + L_S + L_r$	$L_C + L_s$
Accuracy	62.0	-	72.2	73.6	**79.0**
f-mAP	-	51.1	-	77.8	**78.6**
v-mAP	-	48.1	-	79.9	**80.3**

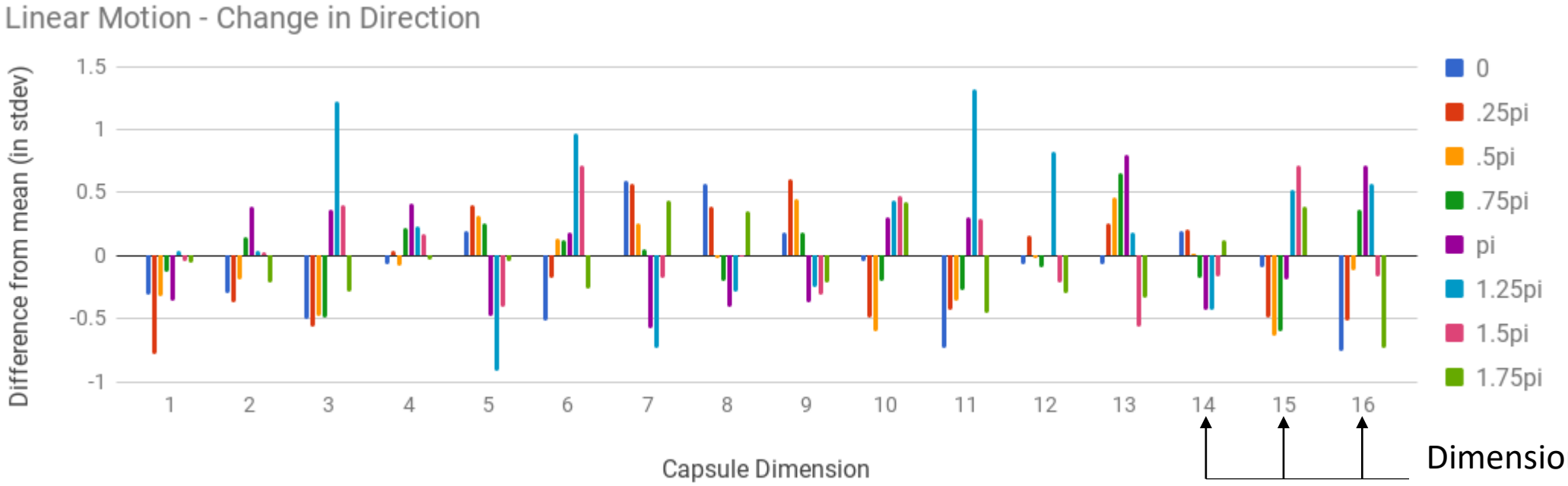
Synthetic Dataset Experiments

- We randomly generate primitive shapes moving in a noisy background
- The shapes have 4 motion types which the network classifies:
  - Linear, circular, a turn, and random
- The primitive shapes vary in shape size, color, speed (constant or accelerating), direction, amount of noise, rotation, and zooming in/out
- Example video clips:



Synthetic Dataset Experiments

- We examine the “linear motion” class capsule’s pose matrix dimensions
- Changes in direction, speed, rotation, zooming in/out, and shape size were predictably reflected in the capsule dimensions



Dimensions correspond with Coordinate Addition